10.48047/jocaaa.2024.33.08.110

DIABETIC TYPE CLASSIFICATION USING SUPERVISED MACHINE LEARNING APPROACHES

SARITA KUMARI[1], DR. AMRITA UPADHAYA[2] ,MONIKA[3]

[1]Research Scholar, Department of Computer Science, Banasthali Vidyapith, Jaipur, Rajasthan, INDIA
Email: saritanaveenkaliraman@gmail.com
[2]Assistant Professor, Department of Computer Science, Banasthali Vidyapith, Jaipur, Rajasthan, INDIA
Email: saiamrita27@gmail.com
[3]Assistant Professor, Department of Computer Science, RPSCET, MAHENDERGARH, INDIA
Email: 001monikayadav@gmail.com

ABSTRACT:

Diabetic Retinopathy (DR) is the leading cause of blindness worldwide and a serious diabetic complication. To prevent vision loss, DR lesion diagnosis and categorization must be done early. DR early detection and treatment can significantly reduce the risk of vision loss. This paper focuses on classifying a sample into diabetic and non-diabetic using a variety of techniques, including Decision Tree, ANN, KNN, SVM, Random Forest, and Gradient Boosting Algorithms. The NCSU Diabetes the data set is pre-processed, and examples are trained and evaluated for accuracy; SVM and ANN achieve over 80% accuracy, demonstrating their potential in diabetes type classification. The PIMA Indians Dataset is used as a reference. The DR's manual diagnosing procedure Ophthalmologists' retina fundus scans take a lot of time, effort, money, and are prone to in contrast to computer-aided diagnosis systems, to misdiagnosis. Machine learning has recently been one of the most widely used methods that has improved performance in several categories, for example. The best classifier for diabetic retinopathy is determined by SVM, Decision. This compares ANN classifiers, Tree, Logistic Regression, and k-Nearest Neighbors paper. Additionally, a study of the existing DR datasets has been conducted. Numerous difficult also covered are topics that need further research. The results of comparing various machine learning algorithms with earlier studies are favourable. This research enhances the diagnosis of diabetic retinopathy by demonstrating the effectiveness of several machine learning classifiers and assisting in the creation of precise and effective computer-aided diagnostic tools for management and early detection.

KEYWORDS : Diabetes, SVM, Decision Trees, KNN, ANN, Gradient, Naive Bayes, Random Forest

10.48047/jocaaa.2024.33.08.110

1. INTRODUCTION

Diabetes is one of the most prevalent lifestyle-related health conditions worldwide, with less than half a billion people currently living with it. The number of people diagnosed with diabetes is expected to continue rising year after year. Type 2 diabetes is the most common form and typically affects older adults who lead sedentary lifestyles and are overweight. Diabetes occurs when the pancreas fails to produce enough insulin or when the body's cells and tissues become resistant to insulin.

Diabetes mellitus is classified into three main types:

1. **Type 1 Diabetes Mellitus**: This type is characterized by the pancreas producing insufficient insulin, a condition also known as Insulin-Dependent Diabetes Mellitus (IDDM). People with type 1 diabetes require daily insulin injections to manage their condition. Early intervention is crucial to prevent complications like diabetic retinopathy, a condition that can lead to blindness. Identifying diabetic retinopathy early can be challenging in many parts of Asia and Africa due to limited access to healthcare. Researchers are exploring artificial intelligence (AI) models that use machine learning to analyze retina fundus images, helping medical professionals detect and classify the stages of diabetic retinopathy, such as Normal, Moderate, and Proliferative Diabetic Retinopathy (PDR).
2. **Type 2 Diabetes Mellitus**: In this form of diabetes, the body becomes resistant to insulin, meaning that the cells do not respond to insulin in the usual manner. Type 2 diabetes, also known as Non-Insulin Dependent Diabetes Mellitus (NIDDM) or Adult-Onset Diabetes, is more common in people who are overweight or lead inactive lifestyles. This is the most prevalent type of diabetes. Diabetic retinopathy (DR), a serious eye condition that can cause vision loss, is often diagnosed in individuals with type 2 diabetes. Manual screening of DR by ophthalmologists can be time-consuming, and thus AI and deep learning (DL) techniques are being used to analyze the different stages of DR, ultimately helping to streamline the diagnostic process.
3. **Gestational Diabetes**: This type of diabetes occurs during pregnancy and typically resolves after childbirth, although women who experience gestational diabetes are at a higher risk of developing type 2 diabetes later in life.

Other, rarer forms of diabetes include Maturity Onset Diabetes of the Young (MODY) and Type 3 diabetes. In healthy individuals, blood glucose levels typically range from 70 to 99 milligrams per deciliter (mg/dL). A fasting blood glucose level greater than 126 mg/dL is indicative of diabetes. Pre-

10.48047/jocaaa.2024.33.08.110

diabetes is a condition where blood glucose levels fall between 100 and 125 mg/dL, increasing the risk of developing type 2 diabetes. Those at higher risk of developing diabetes include individuals who:

- Have pre-diabetes
- Are overweight (BMI ≥ 25)
- Are 45 years or older
- Are physically inactive
- Have a family history of diabetes

Understanding and managing these risk factors is key in preventing and controlling diabetes. Diabetes can have a significant impact on the human body, affecting various organs and systems over time. Here's a breakdown of how it affects different parts of the body:

1. *Blood Vessels and Circulation*

- **Increased Risk of Heart Disease and Stroke**: High blood sugar levels can damage blood vessels, increasing the risk of cardiovascular diseases like heart attack, stroke, and peripheral artery disease.
- **Poor Blood Flow**: Reduced circulation due to blood vessel damage can lead to complications such as leg and foot ulcers, infections, and even amputations in severe cases.

2. *Kidneys (Diabetic Nephropathy)*

- **Kidney Damage**: High blood sugar can damage the blood vessels in the kidneys, impairing their ability to filter waste. This can lead to kidney disease or even kidney failure if not managed properly.
- **Signs of Kidney Problems**: Swelling in the legs, ankles, or feet, and high blood pressure can be early signs of kidney damage.

3. *Eyes (Diabetic Retinopathy)*

- **Vision Problems**: High blood sugar can damage the tiny blood vessels in the retina, leading to diabetic retinopathy, a leading cause of blindness in diabetics.
- **Increased Risk of Cataracts and Glaucoma**: People with diabetes are also at higher risk of developing cataracts and glaucoma, both of which can affect vision.

4. *Nerves (Diabetic Neuropathy)*

- **Nerve Damage**: Chronic high blood sugar levels can damage nerves throughout the body, leading to a condition known as diabetic neuropathy. This can cause pain, tingling, numbness, and weakness, especially in the feet and hands.
- **Digestive Problems**: Nerve damage can also affect the digestive system, leading to issues like nausea, vomiting, diarrhea, or constipation.

5. *Skin*

- **Increased Susceptibility to Infections**: Diabetes weakens the immune system, making the body more susceptible to infections, especially in the skin, urinary tract, and gums.

10.48047/jocaaa.2024.33.08.110

- **Skin Changes**: People with diabetes may experience dry skin, itching, and poor wound healing. Even minor cuts or bruises can take longer to heal.

6. *Mouth and Gums*

- **Gum Disease**: High blood sugar increases the risk of gum disease (periodontitis) by promoting bacterial growth. This can lead to swollen, bleeding gums and tooth loss if left untreated.
- **Dry Mouth**: People with diabetes are also more likely to experience dry mouth, which increases the risk of tooth decay and other oral health problems.

7. *Feet*

- **Poor Circulation and Nerve Damage**: Diabetes can cause poor circulation and nerve damage in the feet, making it more difficult to feel injuries, which increases the risk of infections and ulcers.
- **Amputations**: In severe cases, untreated foot problems or infections can lead to the need for amputations, especially if blood sugar levels are poorly managed.

8. *Hormonal and Metabolic Effects*

- **Blood Sugar Swings**: Poor control of blood sugar can lead to extreme fluctuations, causing episodes of hyperglycemia (high blood sugar) or hypoglycemia (low blood sugar), both of which can be dangerous.
- **Increased Fat Storage**: Insulin resistance (particularly in type 2 diabetes) can lead to the body storing more fat, especially around the abdomen, increasing the risk of obesity-related health issues.

9. *Mental Health*

- **Stress and Anxiety**: Managing diabetes can be stressful, and the constant monitoring of blood sugar levels can cause anxiety for some people.
- **Depression**: People with diabetes are at an increased risk of depression, possibly due to the physical challenges of managing the disease, as well as the impact it has on quality of life.

10.48047/jocaaa.2024.33.08.110

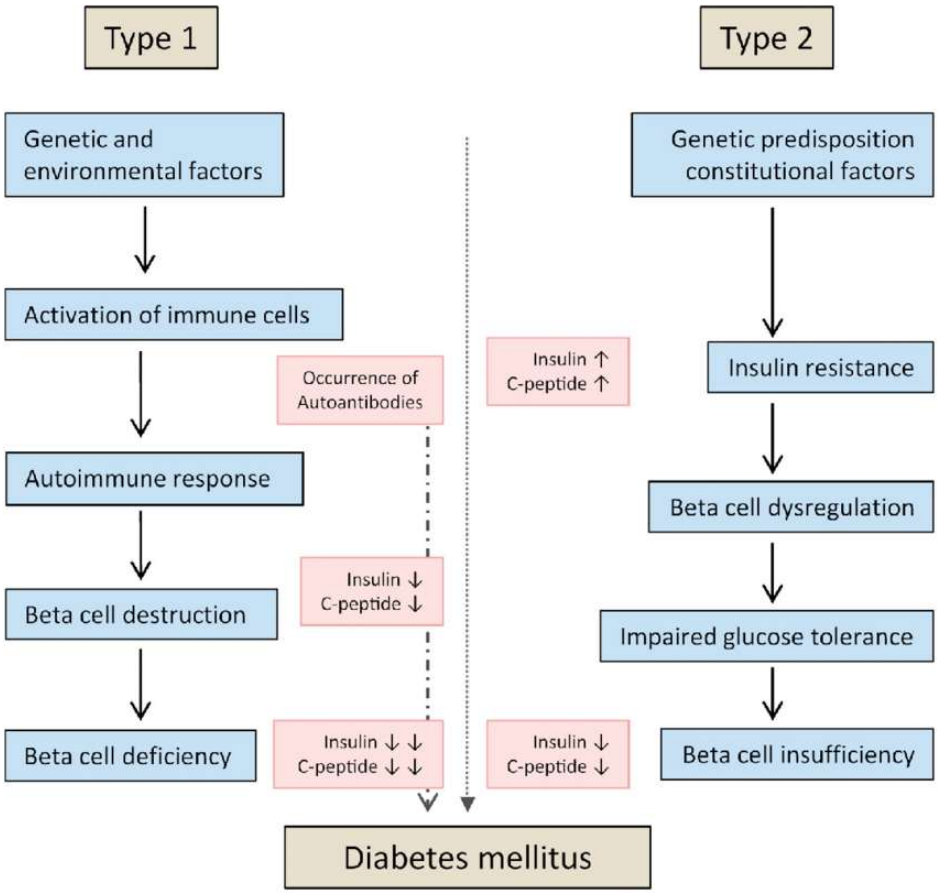

Figure1: Diabetic Type I and Type II flowchart

10. *Reproductive Health*

- **Infertility and Sexual Dysfunction**: Diabetes can affect reproductive health, including erectile dysfunction in men and reduced fertility in women. High blood sugar can affect hormone production and lead to complications in pregnancy, such as gestational diabetes.

11. *Immune System*

- **Weakened Immune Response**: The high blood sugar levels in people with diabetes can weaken the immune system, making it harder for the body to fight off infections.

Diabetes can lead to a wide range of complications if not well-managed, affecting many systems in the body. Proper blood sugar control, a healthy diet, regular physical activity, and consistent monitoring are crucial to minimizing the risk of these complications and maintaining overall health. Regular checkups with healthcare providers can also help identify potential issues early on, allowing for better management and treatment. In figure 1 a simplified flowchart of diabetic is shown that highlights the key differences between Type 1 and Type 2 diabetes, focusing on the causes, onset, and treatment options for each.

Key Differences between Type 1 and Type 2 Diabetes:

10.48047/jocaaa.2024.33.08.110

- **Type 1 Diabetes** is primarily an autoimmune disorder where the body's immune system attacks and destroys the insulin-producing cells in the pancreas. It's more common in younger people and requires lifelong insulin therapy.
- **Type 2 Diabetes** is characterized by insulin resistance, where the body either doesn't respond well to insulin or doesn't produce enough of it. It's more common in older adults and is often related to lifestyle factors, such as being overweight and inactive.

2. LITERATURE REVIEW

In a study from the US, there is 0.5% prevalence of diabetes Type-1 and 8.5% prevalence of diabetes Type-2 among all adults in the US. This paper is an attempt to implement improvised algorithms to tackle the problem. [6] A. Berina et al. (2017) reviewed the classification of diabetes and cardiovascular disease using machine learning methods. Multilayer feed forward neural networks trained using the Levenberg-Marquardt technique predominated in the publications they looked at. However, Naive Bayesian networks are the most used form of BN, and they had the greatest retrospective accuracy rates (99.51for classifying diabetes and CVD, respectively. Additionally, ANN has demonstrated superior results when calculated the mean accuracy of observed networks, suggesting a greater likelihood of obtained correct findings when applied to diabetes and/or CVD classification. N. S. Khan et al. (2017) focused on the machine-learning- based mobile health app for diabetes prediction millions of individuals all around the world suffer with diabetes mellitus, a chronic and lifestyle illness. world succumbs to its effects. While there were applications available for monitoring your diet, blood sugar, medication, and daily activities, measurements of blood sugar, blood pressure, and weight, as well as the distribution of recommendations for diet and exercise in the management of diabetes. There was no known software that was designed specifically to assess the likelihood of complications if you have diabetes. This research aims to do just that by creating a smart Health app. Using AI to predict how likely they are to be successful. whether they are type 1 or type 2 diabetes, pre diabetic, or a healthy medical practitioner, or diagnostic procedures. R. Pal et al. (2017) introduced the diabetic retinopathy as a candidate for use of machine learning algorithms. Clinical data analyzed makes use of machine learning, a subfield of AI, because of its ability to recognise patterns in data and extrapolate those patterns to make inferences or predictions in the face of ambiguity. When dealing with DR, it was crucial to determine whether or if there is a connection between the level of attachment for the patient and the patient's therapeutic outcome. In this research, they evaluate and validate the results of many machine learning techniques for a specific DR data set. I. Kavakiotis et al. (2017) did research on diabetes using machine learning and data mining techniques. The current research aims to conduct a systematic review of the uses of machine learning, data mining techniques, and tools in diabetes research, specifically in

10.48047/jocaaa.2024.33.08.110

the areas of a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, d) Health Care and Management.

Machine learning algorithms from a broad variety of families were used. In general, supervised learning techniques accounted for 85 unsupervised methods, and association rules in particular, accounted for the remaining 15 productive approach was SVM. In terms of data types, clinical datasets were the most common. The publications were chosen because their titles foreshadow the potential for extracting beneficial information that might lead to new hypotheses aimed at a more thorough comprehension of DM. 4 M. A. Sarwar et al. (2018) presented work on the medical diabetes prediction using deep neural networks. In this study, they explore the application of six distinct machine learning algorithms to the field of healthcare predictive analytics. Six different machine learning algorithms are applied to a dataset consisting of medical records from a sample of patients for the aim of this experiment. They examine and contrast the performance and precision of the various algorithms used in this study. This research uses a variety of machine learning methods to determine which algorithm was most effective for diabetes prediction. The purpose of this study was to provide medical professionals with a tool for early diabetes prediction using machine learning methods. Predicting Severe Retinopathy of Prematurity with an Interpretable Machine Learning Approach Using a Generalised Additive Model with Pair wise Interactions (GA2M). Using statistical analysis and logistic regression as a kind of generalised additive model (GAM) with pairwise interaction terms (GA2M), they had looked at the variables that cause serious retinopathy of prematurity. When applying machine learning to clinical data, they address the tradeoff between precision and human understanding. They also validate the experience based judgement of neonatologists about the clinical significance of certain risk variables, such as gender, for RoP prediction. R. Aminah et al. (2019) introduced the machine-learning- supported logical diabetes prediction system. This technique may indicate the health of the organ before any indications of illness occur.

In this study, they used machine learning to develop an iridology or iris-image-based diabetes prediction system. Detection was made easier with the use of machine learning. The system components include hardware and software for capturing and processing images of the eye. The irises were photographed using an iridology camera and analyzer. In order to extract texture features from a picture, the GLCM technique was often used. Diabetic and non- diabetic groups were separated using the k-NN approach. The k-fold cross-validation technique was then used to check and assess the categorization findings using the confusion matrix. There were 16 non-diabetic patients and 11 diabetes subjects in the study. Based on the data, they know that the accuracy is 85.6 percent, the FPR is 11.7 percent, the FNR is 20.4 percent, the specificity is 0.889, and the sensitivity is 0.796. L. Alic et al. (2018) did research on diabetes risk assessment in a healthy population. They use support vector machines with 10 variables that have been shown to be robust predictors of future diabetes in the published literature. The dataset is

10.48047/jocaaa.2024.33.08.110

imbalanced in terms of the class labels, therefore first train the model using 10-fold cross-validation and validate it using a hold-out set. The average validation accuracy across 100 iterations in this investigation was 84.1. The results of this research may be used to better understand who is at risk for getting type 2 diabetes in the future.

3. RELATED WORKS

There are currently very few studies that divide diabetes into Type-1 and Type-2, but in the past, numerous studies have effectively identified whether or not a patient has diabetes. Machine learning techniques can be categorized into different types based on their learning approach and application. Below is an overview of widely used methods:

1. Supervised Learning

In supervised learning, models are trained using labeled datasets, where input-output relationships are predefined. The model learns from historical data to make predictions on new inputs.

- **Regression**: Used for predicting continuous values.
    - Examples: Linear Regression, Polynomial Regression, Support Vector Regression (SVR).
- **Classification**: Assigns data to predefined categories.
    - Examples: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Neural Networks.

2. Unsupervised Learning

Unlike supervised learning, unsupervised learning works with unlabeled data, identifying hidden patterns and structures within datasets.

- **Clustering**: Groups similar data points together.
    - Examples: K-Means, Hierarchical Clustering, DBSCAN.
- **Dimensionality Reduction**: Reduces the number of features while preserving key information.
    - Examples: Principal Component Analysis (PCA), t-SNE, Autoencoders.

3. Semi-Supervised Learning

This approach blends aspects of supervised and unsupervised learning by utilizing a small set of labeled data along with a larger set of unlabeled data to improve model performance. It is commonly applied in areas like image classification and speech recognition.

4. Reinforcement Learning

Reinforcement learning (RL) focuses on decision-making by training an agent to interact with an environment and maximize cumulative rewards.

- **Popular Techniques**:

10.48047/jocaaa.2024.33.08.110

- Q-Learning
- Deep Q Networks (DQN)
- Policy Gradient Methods
- Actor-Critic Algorithms

- **Common Applications**:
  - Robotics
  - Game AI (e.g., AlphaGo, OpenAI Five)
  - Autonomous systems

5. Deep Learning

Deep learning is a specialized subset of machine learning that employs artificial neural networks with multiple layers to process complex data.

- **Convolutional Neural Networks (CNNs)** – Used for tasks such as image recognition and computer vision.
- **Recurrent Neural Networks (RNNs)** – Designed for sequential data like time series forecasting and speech processing.
- **Transformers** (e.g., BERT, GPT) – Primarily used in natural language processing (NLP) for tasks like text generation and translation.

6. Ensemble Learning

Ensemble learning improves prediction accuracy by combining multiple models to enhance overall performance.

- **Bagging**: Uses multiple weak models to reduce variance (e.g., Random Forest).
- **Boosting**: Iteratively enhances weak models to strengthen performance (e.g., XGBoost, AdaBoost).
- **Stacking**: Merges different models to create a stronger predictive framework.

These machine learning techniques serve various industries, from healthcare and finance to automation and artificial intelligence applications.

4. SUPERVISED LEARNING

Supervised learning is well-suited for diabetes prediction as it enables models to analyze past medical data with known outcomes (such as diabetic or non-diabetic cases). Popular algorithms for this purpose include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and Neural Networks.

4.1. Logistic Regression

10.48047/jocaaa.2024.33.08.110

Logistic regression, a method in data analysis that is similar to a supervised learning algorithm, predicting outcomes with limited possibilities like yes or no. It utilizes the logistic function to estimate probabilities and holds significance in AI/ML for actionable insights. If the output of the Sigmoid function is greater than 0.5, the output will be 1, else it will be 0, this technique is extremely helpful in classification kind of problems [7].

$$S(x) = 1/(1 + e^{-x})$$

4.2. K-Nearest Neighbor Algorithm

The k-nearest neighbors (KNN) algorithm is a non-parametric supervised learning classifier used for classification or regression based on proximity. It operates on the idea that similar data points are close to each other and it does on the basis of calculating the Euclidean distance. In classification, a class label is assigned through a majority vote, where the most frequently represented label near a data point is used. While technically"plurality voting," the term"majority vote" is commonly used. This method is effective for multiple classes, as it doesn't strictly require a majority of over half when dealing with more than two categories [8].

4.3. Support Vector Machine

Support Vector Machine (SVM) is a powerful supervised machine learning algorithm designed for both classification and regression tasks. It identifies an optimal hyper plane that effectively separates data points of different classes in a high-dimensional space. The primary objective is to find the hyper plane with the maximum margin, representing the maximum distance between classes. SVM is particularly suitable for scenarios with smaller datasets, excelling in cases of complex classification problems [9]. It categorizes data points based on their proximity to the hyper plane, with support vectors playing a crucial role in defining the separating line. SVM's efficacy is notable in cases of linearly separable data, and it offers versatility through various kernel functions for handling non-linear datasets [10].

4.4. Decision Tree

Decision Trees (DTs) are non-parametric supervised learning methods used for classification and regression [11]. They create models by learning simple decision rules from data features, resulting in piecewise constant approximations. Decision Trees can handle multi-output problems and offer statistical test validation. However, they may suffer from over fitting, instability and discontinuous predictions [12].

4.5. Random Forest

Random Forest is a versatile machine learning algorithm that combines outputs from multiple decision trees, reducing the risk of over fitting. Decision trees, the algorithm's building blocks, use questions to

10.48047/jocaaa.2024.33.08.110

split data into categories. Random Forest employs ensemble methods like bagging and feature randomness, ensuring uncorrelated trees. With three key hyper parameters, it creates an ensemble of decision trees trained on bootstrap samples, cross-validated with an out-of-bag sample [4].

4.6. Naive Bayes

Naive Bayes is a powerful classification algorithm widely used in machine learning, particularly for text classification. The algorithm is based on Bayes' Theorem, which calculates the probability of a hypothesis given the data. In classification, it assigns a class label to a new data instance based on the highest posterior probability. The Naive Bayes assumption simplifies computations by assuming conditional independence of features, making it particularly effective for categorical or binary input values [13].

4.7. Gradient Boosting

Machines Light GBM is a powerful gradient boosting framework in machine learning known for its speed and efficiency in handling large datasets. Unlike traditional algorithms, Light GBM adopts a leaf-wise growth strategy, selecting the leaf with the highest loss, resulting in faster and more accurate computations [14].

4.8. Ensemble method

Ensemble methods represent a sophisticated approach in machine learning wherein multiple models, such as decision trees or neural networks, are amalgamated to enhance predictive accuracy. Each constituent model within the ensemble contributes its unique strengths, thereby compensating for individual weaknesses. By aggregating the predictions of diverse models, ensembles achieve superior performance and robustness compared to individual models [15].

5. METHODOLOGY

The methodology for predicting diabetes using supervised learning typically involves several key steps, from data collection to model evaluation.

5.1. Datasets

Description [16] To conduct this study we have used 2 datasets, North Carolina State University's Diabetes dataset (NCSU) d & benchmark PIMA Indian dataset to validate our model's performance.

5.2. Design & Implementation

To facilitate our research study, Python programming language has been used in order for us to implement various algorithms such as Logistic Regression, KNN, SVM, ANN, Ensemble, and Various Gradient Boosting Algorithms. Sci-Kit Learn library was used to implement the algorithms. Parallels

10.48047/jocaaa.2024.33.08.110

Module was used to speed up the computational training of various models with features combinations. The Datasets were cleaned & pre-processed, after pre-processing; they were trained over all possible features combinations to select the best feature combinations resulting in highest accuracy for every Machine Learning Model over both the datasets. This Methodology resulted in higher accuracy for both datasets in comparison to other papers. As shown in figure 2 the adopted methodology for the related work how we are getting our required results by following these steps Understanding, pre-processing, Analysis, Implementation and Results Analysis etc.

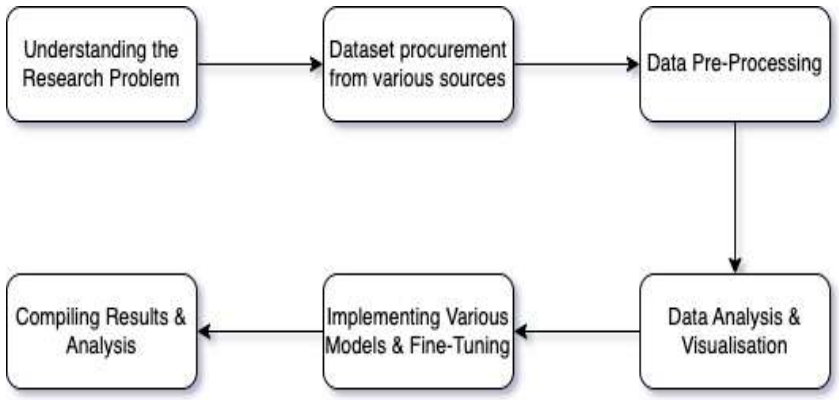

Figure 2: Methodology flow chart

Table 1: Model Advanced Performance Metrics

Model	F-Score	MCC	10-foldCV	Kappa	AUC	CVVariance
LR	0.730769	0.605386	0.766934	0.596255	0.785920	0.000556
KNN	0.693069	0.561763	0.722710	0.548173	0.760057	0.001799
RF	0.693069	0.561763	0.747437	0.548173	0.760057	0.001474
DT	0.600000	0.380171	0.688688	0.378805	0.685524	0.002424
ANN	0.716981	0.576411	0.759193	0.570472	0.775503	0.002201
SVM	0.700000	0.577201	0.752632	0.561170	0.765266	0.000526
NB	0.673913	0.587880	0.748753	0.548122	0.751616	0.000954
LGBM	0.684211	0.498928	0.738329	0.498734	0.747665	0.003543
XGB	0.694915	0.505721	0.734433	0.505530	0.754490	0.001160
CAT	0.725664	0.567365	0.759193	0.566866	0.780532	0.002710
Ensemble	0.750000	0.608386	0.755161	0.607429	0.799969	0.002757

Table2: Features comparison of existing Datasets

10.48047/jocaaa.2024.33.08.110

Algorithm	NCSU Dataset Features	PIMA Dataset Features
LR	PR,GL,BP,BMI	PR,GL,BP,BMI
KNN	GL,Age,OB	PR,GL,BP,ST,Age
RF	PR,GL,Age,BP,ST,IN,DPF,VB HDL, IN,DPF,VB,HDL	IN,DPF,VB,HDL
DT	GL,BMI,OB,HDL	PR,GL,BP,ST,DPF,Age
ANN	GL,IN,BMI,DPF,VB,HDL	PR,GL,DPF
SVM	GL,BMI,DPF,PP,VB,OB,HDL	PR,GL,DPF,BP,IN,Age
NB	GL,BP,VB,OB	PR,GL,BP,BMI,Age
LGBM	PR,Age,GL,BP,ST,IN,BMI,DPF PP,VB, OB,Smoker, HDL	PR,GL,BP,BMI, DPF
XGB	PR,Age,GL,BP,ST,IN,BMI,DPF PP,VB, OB,Smoker, HDL	PR,GL,BP, BMI, DPF, ST, IN
CATBGM	PR,Age,GL,BP,ST,IN,BMI,DPF PP,VB, OB,Smoker, HDL	PR,GL,DPF
Ensemble	PR,Age,GL,BP,ST,IN,BMI,DPF PP,VB, OB,Smoker, HDL	PR,Age,GL,BP,ST,IN,DPF

6. RESULTS AND DISCUSSION

After extensive computational analysis, the features listed in the following table emerged as the optimal choices for each algorithm. These features were identified through rigorous experimentation and evaluation, encompassing various machine learning models and datasets. Here the Table1 demonstrate various algorithms with their different datasets used on PIMA dataset and CSMU dataset.

7. CONCLUSION

Diabetes is one of the major health issues that is plag using the world, by implementing this methodology we have achieved a high accuracy of 81% on the benchmark PIMA dataset and around 83% in the NCSU dataset, while retaining good F-1 scores and other parameters, there have been previous works in this field but nobody utilized the technique of features optimization to find out which features work the best for a specific dataset. In our prospective investigations, we endeavor to extend the application of this methodology across an array of datasets, thereby seeking to refine and bolster the accuracy of our analytical techniques. Through the systematic exploration of diverse data sources, we aim to gain deeper insights into the efficacy and versatility of our approach in various research domains. Our commitment to methodological refinement and iterative

10.48047/jocaaa.2024.33.08.110

optimization underscores our dedication to advancing predictive analytics in the realm of health sciences [17].

Table3: Model Basic Performance Metrics

Model	Accuracy	Error	Sensitivity	Specificity
LR	0.818182	0.181818	0.655172	0.916667
KNN	0.798701	0.201299	0.603448	0.916667
RF	0.798701	0.201299	0.603448	0.916667
DT	0.714286	0.285714	0.568966	0.802083
ANN	0.805195	0.194805	0.655172	0.895833
SVM	0.805195	0.194805	0.603448	0.927083
NB	0.805195	0.194805	0.534483	0.968750
LGBM	0.766234	0.233766	0.672414	0.822917
XGB	0.766234	0.233766	0.706897	0.802083
CAT	0.798701	0.201299	0.706897	0.854167
Ensemble	0.818182	0.181818	0.724138	0.875000

REFRENCES

[1] Diabetesdetectionusingmachinelearningclassificationmethods. URLehttp://dx.doi.org/10.1109/ICIT52682.2021.9491788

[2] Diabetespredictionusingdifferentmachinelearningclassifiers.URLewww.ijtrd.com

[3] Typesofdiabetes, https://www.diabetes.org.uk/diabetes-the-basics/types-of-diabetes.

[4] Randomforest. URLhttps://www.ibm.com/topics/random-forest

[5] Diabetesriskfactors, https://www.cdc.gov/diabetes/basics/risk-factors.html.

[6] G.Xu,Prevalenceofdiagnosedtype1andtype2diabetesamongusadultsin2016an d 2017:populationbasedstudy,Britishmedicaljournal(2018).

[7] Logisticregression.URLhttps://aws.amazon.com/what-is/logistic-regression/

[8] K-nearestneighborsalgorithm. URL"https://www.ibm.com/topics/knn"

[9] Diabetespredictionusingmachinelearningandexplainableaitechniques. URL10.1049/htl2.12039

[10] Svm. URL"https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-comp

[11] Diabeticretinopathydetectionusingvgg-

10.48047/jocaaa.2024.33.08.110

ninadeeplearningarchitecture.
URLhttps://creativecommons.org/licenses/by/4.0/

[12] Decision trees.URLhttps://scikit-learn.org/stable/modules/tree.html

[13] Naivebayes.URLhttps://machinelearningmastery.com/naive-bayes-for-machine-learning/

[14] Lightgbmalgorithm.
URLhttps://www.analyticssteps.com/blogs/what-light-gbm-algorithm-how-use-it

[15] Ensemblemethodsinmachinelearning.
URLhttps://dl.acm.org/doi/10.5555/648054.743935

[16] Predictionofdiabetesempoweredwithfusedmachinelearning.
URLehttps://creativecommons.org/licenses/by/4.0/

[17] S. M. S. H. Z. Saquib, Diabetic retinopathy detection using deep learning, 2020 Inter- national Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (2020).
URL110.1109/ICSTCEE49637.2020.9277506